Introducción a Stata 11.0

USCO

Temas a Desarrollar

- Introducción a la interface
- Sintaxis
- Comandos básicos.
- Organización de un proyecto en archivo log y do.
- Introducción a regresiones básicas

Introducción a Stata

- STATA es un sistema que permite la gestión de base de datos y la realización de cálculos estadísticos y econométricos.
- La gran ventaja de STATA es que se basa en un lenguaje de programación que respeta una sintaxis. Quienes sepan principios de programación podrán asimilar a STATA como un sistema de programación de alto nivel con algunos aspectos similares a Pascal, C, o Basic.
- La otra ventaja es que existen comandos predefinidos para realizar tareas muy complejas.
- Por otra parte, STATA admite la creación de nuevos comandos por parte del usuario, lo que le otorga amplias posibilidades de expansión.

Interface del Stata

- Ventana “Review”: muestra el historial de comandos recientemente utilizados.
- Ventana “Variables”: expone las variables que comprenden el *dataset* actualmente en memoria.
- Ventana “Results”: es la ventana más grande color negro. Muestra los resultados obtenidos de la aplicación de los comandos.
- Ventana “Command”: en esta ventana el usuario introduce comandos mediante el teclado (contiene un *prompt).*

Sintaxis y comandos básicos

STATA trabaja mediante la especificación por parte del usuario de órdenes que se denominan comandos. Los comandos conforman un lenguaje de comunicación el programa, por lo que existe una determinada sintaxis que debe ser respetada.

La sintaxis de comandos tiene la siguiente estructura general:

[by varlist:] **comando** **[varlist] [=exp] [if exp] [in range] [weight] [, opciones]**

Los corchetes indican elementos opcionales. Todos los elementos que acompañan a un comando son opciones, de hecho existen comandos que comprenden sólo una palabra.

Abrir el archivo base1.dta, para desarrollar algunos ejemplos en donde se utilizan los comandos (Este dataset contiene datos de 74 automóviles incluyendo variables como precio, peso, largo, marca, consumo, etc.)

Sintaxis: comando sin opciones

summarize: sum

- Si se utiliza solo, realiza una descripción estadística de todas las variables que componen el dataset.

- **Sintaxis: [varlist]**

El anterior comando puede usarse para obtener la descripción estadística de sólo una variable

summarize price

ó de más variables

summarize price weight length

ó de una lista de variables (en este caso todas las variables que están entre price y trunk)

Summarize price-trunk

Ó de las variables que empiecen con la letra p (en este caso es sólo price)

Summarize P*

- **Sintaxis: [if exp]**

La aplicación del comando puede restringirse a observaciones que cumplen con ciertas restricciones de alguna/s variable/s. Esto se realiza mediante la utilización del “si condicional” (if en inglés). Por ejemplo:

summarize length if price > 6000

Esta expresión calcula la descripción estadística de la variable length de aquellos autos cuyo precio es superior a 6000.

- **Sintaxis: [in range]**

Permite aplicar el comando a un rango de observaciones, de acuerdo al orden del dataset. Por ejemplo:

summarize price in 1/10

Permite obtener la descripción estadística de las 10 primeras observaciones del dataset.

summarize price in 5/12

Aplica el comando a las observaciones entre la 5ta y la 12da.

summarize price in -10/-1

Aplica el comando a las últimas 10 observaciones (de la 65 a la 74).

La aplicación de [in range] depende del ordenamiento del dataset. Los comandos para ordenar un dataset son sort y gsort.

sort price

El comando sort permite ordenar sólo de manera ascendente de acuerdo a la variable que se especifica.

gsort – price

Ordena de manera descendente según el precio del auto.

gsort + price

Ordena de manera ascendente según el precio del auto.

- **Sintaxis: [= exp]**

Se utiliza generalmente con sólo dos comandos: **generate** **y** **replace.**

generate nueva = 0

Este comando permite crear una nueva variable en el dataset. Es requisito indicar los valores iniciales que va a tener la nueva variable en cada observación. En este caso la variable se llama nueva y tiene valor 0 en todas las observaciones.

También puede crearse una variable nueva inciándola con valores surgidos de una operación algebraica basada en otras variables. Por ejemplo:

generate price2 = price*2

Este comando crea la variable price2 igual a los valores de price multiplicados por 2 en cada observación.

Se puede compara las dos variales usando **list.**

Otro comando que usa [=exp] es **replace**. Este comando permite reemplazar valores de una variable ya creada. Por ejemplo:

replace price2 = 0 if price < 6000

Este comando reemplazó por un cero en price2 en aquellas observaciones que tenían price menor que 6000.

replace price2= 1 if price2>=19000 & price2<30000

- **Sintaxis: [by varlist:]**

Permite aplicar el comando por grupos de observaciones definidos por los distintos valores de la variable indicada en varlist. Por ejemplo:

sort foreign

by foreign: summarize price

Es requisito ordenar el dataset por la variable que se va a usar en el [by varlist:]. En este caso se obtuvo la descripción estadística para los autos domésticos e importados.

by foreign: tab price

Otros comandos básicos

Para tabular cantidades de observaciones según distintos valores de una variable, tenemos el comando **tabulate. Por ejemplo:**

tabulate foreign

Permite conocer cuántas observaciones hay con los diferentes valores que tiene la variable foreign (Domestic ó Foreign).

El tabulate se puede combinar con otros comando u otras variables. Por ejemplo:

tabulate foreign rep78

Permite conocer cuántas observaciones hay con las diferentes combinaciones de valores entre las variables foreign y rep78.

Obviamente, cualquiera de estos comandos acepta la introducción de condiciones lógicas, como por ejemplo:

tabulate foreign rep78 if price > 3000

El comando **tabstat es más avanzado para obtener la descripción estadística de** variables. El uso es el siguiente:

tabstat price, statistics(mean)

Obtenemos la media de la variable price Pero también podemos obtener muchos otros estadísticos descriptivos, como por ejemplo

tabstat price, statistics(mean sum max min sd variance p10 median)

…y de más de una variable

tabstat price length weight, statistics(mean sum max min sd variance p10 median)

Memoria

STATA trabaja generalmente de una manera particular con los dataset. Todo el dataset es almacenado en la memoria RAM de la PC. Esto significa ventajas y desventajas:

- Ventajas: la ejecución de comandos es muy rápida, dado que no se accede al disco rígido.
- Desventajas: el tamaño de los datasets que pueden procesarse se encuentra limitado por la cantidad de memoria RAM de la PC.

El comando para conocer el estado del uso de la memoria es:

Memory

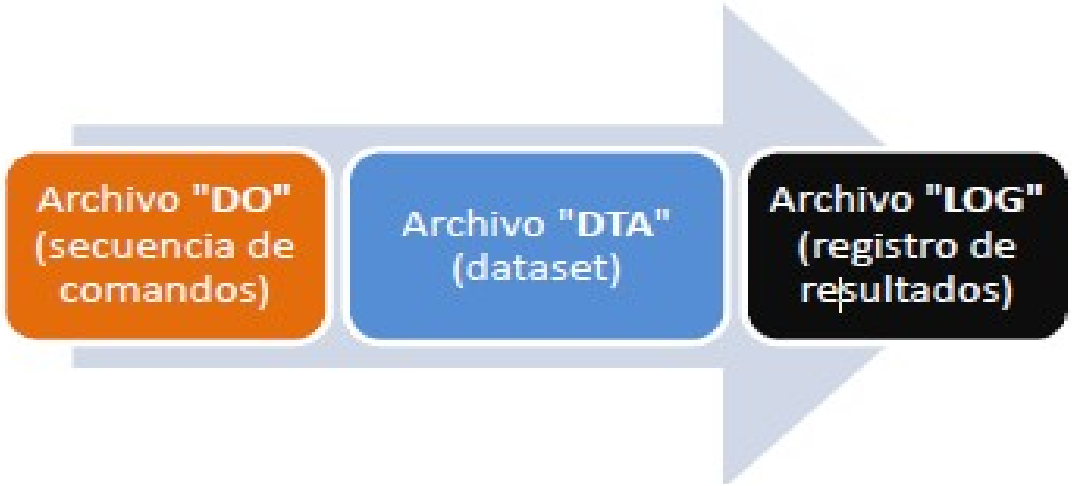
Para ampliar la capacidad de la memoria de Stata, se usa el comando:

set memory 500m

proyecto en archivos “DO” y “LOG”

- Hasta ahora la interacción con Stata ha sido mediante el tipeo de comandos en la ventana “**Commands”. A partir de ahora trabajaremos creando archivos de texto** que contienen toda la secuencia de comandos. Esto es un archivo “**DO”.**
- Adicionalmente, los resultados de nuestro proyecto los almacenaremos en un archivo de texto llamado “**LOG”.**

El síntesis, el conjunto de archivos que utilizaremos en un proyecto serán:



¿Cómo creamos un archivo DO?

Un archivo DO es de tipo “texto plano”.

Tenemos dos alternativas:

1. Utilizar el editor de textos que tiene incluido Stata.
2. Utilizar un editor de textos externo (Editplus, Textpad, Crimson).

Editor de textos incluido en Stata (Do-file Editor)

Para acceder a esta funcionalidad el comando es:

doedit

Archivos Do

Utilizando el dataset **base1.dta escribiremos nuestro primer archivo DO mediante el** editor de textos de STATA. Para ello realizamos las siguientes acciones en el archivo DO:

use "C:\Users\mi\Desktop\seminario economia\base1.dta", clear
describe
summarize price
exit

Nota: pueden incorporarse comentarios dentro del archivo DO de la siguiente manera:

/* Este es un comentario */

También es un comentario una línea iniciado con asterisco *

Archivos Log

Los resultados que se registran en la ventana Stata result pueden ser almacenados en un archivo de texto de extensión **.log**

El código a agregar para obtener un archivo LOG es el siguiente:

```
use "C:\Users\mi\Desktop\seminario economia\
    base1.dta", clear
capture log close
log using "C:\Users\mi\Desktop\seminario economia\base1.log",
    text replace
describe
summarize price
log close
```

Introducción a regresiones básicas

El comando en Stata para estimar un modelo de Mínimos Cuadrados Ordinarios (MCO) es **regress. El siguiente código estima dos modelos de MCO después de** definir el dataset como una serie de tiempo con periodicidad diaria.

```
regress y x
regress y x w
```

El primer modelo estimado es $y_t = \beta_0 + \beta_1 x_t + u_t$

El segundo modelo estimado es $y_t = \beta_0 + \beta_1 x_t + \beta_2 w_t + u_t$

La sintaxis del comando **regress indica que la primer variable representa el** término dependiente del modelo, y las siguientes variables son los términos independientes.

Salida tradicional de una regresión

```
(1)  Source |   (2)  SS      (6) df        MS     (10)        Number of obs =     1500  (14)
-------------+------------------------------------           F(  1,  1498) = 7046.83  (15)
       Model |(3)77120.9915 (7)   1  77120.9915 (11)         Prob > F      =  0.0000  (16)
    Residual |(4)16394.2095(8)1498  10.9440651 (12)          R-squared     =  0.8247  (17)
-------------+------------------------------------           Adj R-squared =  0.8246  (18)
       Total | (5) 93515.201(9)1499  62.3850574 (13)         Root MSE      =  3.3082  (19)

------------------------------------------------------------------------------
(20)      y |  (23) Coef.(24)Std. Err.  (25) t (26) P>|t|  (27) [95% Conf. Interval]
-------------+----------------------------------------------------------------
(21)      x |   2.457744   .0292779    83.95   0.000     2.400314    2.515174
(22)  _cons |   10.63412   .1714624    62.02   0.000     10.29778    10.97045
------------------------------------------------------------------------------
```

Referencias **(..)**

(1) Fuentes de la varianza: en esta parte se muestra la descomposición de la varianza. La varianza total (Total) se descompone en la varianza explicada por el modelo (es decir, por la/s variable/s independiente/s) (Model) y en la varianza no explicada por el modelo (Residual). Notemos que Total = Model + Residual.

la varianza no explicada por el modelo (Residual). Notemos que **Total = Model + Residual**.

(2) **Suma de cuadrados** (Sums os Squares) asociadas a las dos fuentes de la varianza. Estas son medidas de variabilidad respecto de la media.

(3) **Suma de cuadrados explicada:** $ESS = \sum \hat{y}_t^2$

(4) **Suma de cuadrados residual:** $RSS = \sum e_t^2$

(5) **Suma de cuadrados totales:** $TSS = \sum y_t^2$

(6) **Grados de libertad** (degrees of freedom): estos son los grados de libertad asociados a las fuentes de la varianza.

(7) **Grados de libertad del modelo:** son $k - 1$, donde k es el número de variables explicativas (incluyendo la constante).

(8) **Grados de libertad del residuo:** son los grados de libertad totales menos los grados de libertad del modelo: (8)=(9)-(7)

$$F=((SCRr-SCRnr)/q)/(SRCnr/(n-k-1))$$

(9) **Grados de libertad totales:** la varianza total tiene **N – 1** grados de libertad, donde N es el número de observaciones.

(10) **Mean Squares**: MS=SS/df, es decir, es igual a la suma de cuadrados dividida por los grados de libertad respectivos. Con estos datos se puede construir el estadístico F (ver (15)).

(11) **Mean Squares del modelo:** (3) / (7)

(12) **Mean Squares del residuo:** (4) / (8)

(13) **Mean Squares total:** (5) / (9)

(14) **Número de observaciones:** N

(15) **Estadístico F de significatividad global:** $F = \frac{ESS/(k-1)}{RSS/(N-k)} = \frac{(3)/(7)}{(4)/(8)} = \frac{(11)}{(12)}$

(16) **Valor *p* del test de significatividad global**

(17) **R cuadrado:** $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{(3)}{(5)} = 1 - \frac{(4)}{(5)}$

(18) **R cuadrado ajustado:** $\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k} = 1 - (1 - (17))\frac{(14)-k}{(14)-1}$

(19) **Root Mean Squared Error** (o Residual): es el desvío estándar del término de error.

(20) **Variable explicada**

(21) **Variables explicativas:** en este caso el modelo sólo tiene una variable explicativa pero si tuviera más, cada una se mostraría en una fila aparte.

(22) **Constante**

(23) **Vector de los coeficientes estimados:** $\hat{\beta}$

(24) **Desvío estándar de los coeficientes estimados:** $\sigma_{\hat{\beta}} = \sqrt{V(\hat{\beta})}$

(25) **Estadístico t:** $t = \hat{\beta} / \sqrt{V(\hat{\beta})}$

(26) **Valor p de significatividad individual** (con dos colas)